ERA Ranking Representability: The Missing Link Between Ordinal Regression and Multi-class Classification

Willem Waegeman and Bernard De Baets
Department of Mathematical Modelling, Statistics and Bioinformatics
Ghent University, Coupure links 653, B-9000 Ghent, Belgium
Email: forname.surname@ugent.be

***Abstract*—Can a multi-class classification model in some situations be simplified to an ordinal regression model without sacrificing performance? We try to answer this question from a theoretical point of view for one-versus-one multi-class ensembles. To that end, sufficient conditions are derived for which a one-versus-one ensemble becomes ranking representable, i.e. conditions for which the ensemble can be reduced to a ranking or ordinal regression model such that a similar performance on training data is measured. As performance measure, we use the area under the ROC curve (AUC) and its reformulation in terms of graphs. For the three-class case, this results in a new type of cycle transitivity for pairwise AUCs that can be verified by solving an integer quadratic program. Moreover, solving this integer quadratic program can be avoided, since its solution converges for an infinite data sample to a simple form, resulting in a deviation bound that becomes tighter with increasing sample size.**

***Keywords*-one-versus-one multi-class classification, ordinal regression, ranking representability, ROC analysis, cycle transitivity, graph theory, learning theory**

I. Introduction

Multi-class classification and ordinal regression can be seen as two closely related machine learning settings that share many properties. Multi-class classification refers to the supervised learning problem of inferring a predictive model capable of classifying data into a finite number of classes. This simply means that the model predicts for new data instances an output (also called label or response variable) that takes values in a finite unordered set (for example, class labels red, green, blue). Ordinal regression considers a slightly different setting. Labels here come from a finite ordered set, in which the order naturally follows from the semantics of the classes (for example, class labels bad, moderate, good). As a specific case of preference learning, ordinal regression problems typically arise in situations where humans are involved in the data generation process, like human experts or internet users expressing preferences on objects w.r.t. characteristics such as quality, beauty, appropriateness, etc.

So, the different semantics of the data respectively result in the absence or presence of an order relation on the classes in multi-class classification or ordinal regression. Owing to this important interpretation of the classes, substantially different methods have been proposed in the past for the two types of learning problems. Briefly summarized, the absence or presence of an order relation leads to two main differences in assumptions:

1) Firstly, both models typically differ in the type of performance measure they optimize. If an order relation on the classes can be assumed, then a performance measure that takes this order into account must be utilized, both for optimization and evaluation. For example, in ordinal regression, misclassifying an object of class "bad" into class "good" must typically lead to a higher loss than misclassifying the same object into class "moderate".
2) Secondly, the absence or presence of an order relation on the classes gives rise to a different model structure for the two types of problems. The model structure of multi-class classification methods typically consists of an ensemble of binary classifiers, such as one-versus-one [1], [2] and one-versus-all [3] ensembles, while typically only one global model is considered in ordinal regression. Moreover, this global model always consists of an underlying latent variable that reflects the order on the classes. Let $\mathcal{X}$ denote the set of data objects, then this latent variable serves as a ranking function $f : \mathcal{X} \to \mathbb{R}$ that defines a total order on the data objects. The final decision rule is then in the end obtained by placing a number of thresholds on the ranking function. This is for example the case in traditional statistical ordinal regression algorithms [4], [5] and kernel-based methods [6], [7].

Several authors [8]–[10] empirically analyzed in recent work the relationship between multi-class classification and ordinal regression, in which they primarily aim to improve ordinal regression algorithms by using ideas from multi-class classification, without considering an underlying ranking function. Conversely, the motivation of this article is to improve multi-class classification algorithms by using techniques from ordinal regression. Moreover, we will mainly focus on the theoretical connections between both problem settings, and to establish such a connection, we will take the ranking function that characterizes ordinal regression

models as starting point. In this context, expected ranking accuracy (ERA) is a ranking-based performance measure that has recently been introduced for bipartite ranking [11] and further extended to ordinal regression [12]. Expected ranking accuracy can be easily considered too in multi-class classification, especially for one-versus-one ensembles, where the ensemble contains a set of pairwise bipartite ranking functions (i.e. one bipartite ranking function for each pair of classes). By using concepts from receiver operator characteristics (ROC) analysis, graph theory, decision theory and preference modeling, we will show that transitivity properties of the reciprocal relation generated by expected ranking accuracy result in a connection between multi-class classification and ordinal regression models.

Roughly speaking, we will investigate the conditions for which a one-versus-one ensemble, containing a set of bipartite ranking functions, can be reduced to an ordinal regression model with only one underlying ranking function, such that both models obtain an identical performance in terms of expected ranking accuracy. We will further refer to this property as ERA ranking representability of a one-versus-one ensemble. ERA ranking representability can be interpreted as a natural extension to the infinite sample case of AUC ranking representability, as previously introduced in [13]. It is well known that the area under the ROC curve (AUC) forms an unbiased estimator of the expected ranking accuracy on a finite dataset.

II. Ranking Representability

Let us as an introductory example in a multi-class classification setting consider the following hypothetical three-class dataset that contains six objects of each class:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y_i	c_1	c_1	c_1	c_1	c_1	c_1	c_2	c_2	c_2	c_2	c_2	c_2	c_3	c_3	c_3	c_3	c_3	c_3

given a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ of $N = 18$ input-output pairs. We have for simplicity assigned the indices in such a way that pairwise AUCs can be computed easily for a given ranking. Remark that the AUC simply computes the fraction of (lower class, higher class) couples that are correctly ranked by the classifier. Let us suppose that the following triplet of bipartite ranking functions is statistically inferred by a one-versus-one ensemble for this small toy problem:

	i											
ranking for f_{12}	7	8	1	2	9	3	4	5	6	10	11	12
ranking for f_{23}	13	7	14	8	9	10	11	12	15	16	17	18
ranking for f_{13}	13	1	2	3	14	15	16	17	18	4	5	6

So, from left to right, the numbers represent the ranking of the indices of the data objects, respectively obtained with the ranking functions f_{12}, f_{23} and f_{13}. For the pairwise AUCs we find:

$$\begin{aligned} \widehat{A}_{12}(f_{12}, D) &= 20/36\,, \\ \widehat{A}_{23}(f_{23}, D) &= 25/36\,, \\ \widehat{A}_{13}(f_{13}, D) &= 15/36\,. \end{aligned} \tag{1}$$

In other words, one finds for instance that 20 of the 36 couples are correctly ranked by the ranking function f_{12}: object number 1 is ranked before four objects of class $\mathcal{C}_2$, as well as object number 2, object number 3 is ranked before three objects of class $\mathcal{C}_2$, and so on. A more formal definition of the AUC will be given in Section 3.

In this example, the triplet of bipartite rankings can still be replaced in different ways by a single ranking of the whole data set such that the same pairwise AUCs are measured, for example

	i																	
global f	13	1	2	3	7	8	9	10	11	14	15	16	17	18	4	5	12	6

is such a ranking that results in the same pairwise AUCs. Verification of AUC ranking representability is much more difficult for larger datasets, since enumerating all global rankings is then computationally infeasible. However, in [13] we have shown that AUC ranking representability is strongly linked with a specific type of transitivity that has been called AUC transitivity for this reason.

III. Expected ranking accuracy

In the last decade, the problem of ranking, i.e., statistically inferring the parameters of a ranking function $f : \mathcal{X} \to \mathbb{R}$ from a finite data set, has grown out to an active and widespread research field that covers applications like information retrieval, marketing, financial forecasting and more traditional decision making problems (see e.g. [14]–[17]). We in particular focus on pairwise bipartite ranking in a multi-class setting. Such a setting basically implies that one aims to construct a statistical model that describes the relationship between data objects $\mathbf{x} \in \mathcal{X}$ on the one hand and a (usually small) unordered set of r classes $\mathcal{Y} = \{\mathcal{C}_1, ..., \mathcal{C}_r\}$ on the other hand. Although different methods have been proposed for extending binary classification algorithms ($r = 2$) to multi-class classification ($r > 2$), the pairwise approach [1], [2] has been especially popular due to its simplicity, good performance and generality. This approach in essence fits a binary classifier to the data for each pair of classes. It is for this reason also called a one-versus-one classification scheme. Since many binary classification methods like logistic regression, linear discriminant analysis, neural networks and support vector machines construct internally a latent continuous variable, a set $\overline{\mathcal{F}}$ of bipartite ranking functions $f_{kl} : \mathcal{X} \to \mathbb{R}$ is in this way obtained, with $1 \leq k < l \leq r$. These ranking functions can then be further used to generate multi-class probability estimates [18]. For a given data set, the ranking returned by each of the pairwise ranking functions is called bipartite, because it can be visualized by means of a bipartite graph in which the two subsets of nodes correspond to the data instances of the two classes and edges indicate the ranking order of two objects of different classes.

Ranking can be considered somewhere in the middle between pure discriminative modeling (we want good class predictions) and probability estimation (we want good predictions of class-conditional probabilities). The difference between both approaches is in the first place characterized by the type of loss or error function that is optimized. To this end, [11] introduced for ranking the concept of expected ranking accuracy as loss function. In a multi-class setting it can be formally introduced as follows.

Definition 3.1: Let $\mathcal{D}_j$ represent the conditional distribution over $\mathcal{X}$ given that the data object belongs to class $\mathcal{C}_j$ with $j = 1, ..., r$. For a set $\overline{\mathcal{F}} = \{f_{kl} \mid 1 \leq k < l \leq r\}$ of bipartite ranking functions, we define the pairwise expected ranking accuracy between classes $\mathcal{C}_k$ and $\mathcal{C}_l$ for the ranking function f_{kl} as

$$\begin{aligned} A_{kl}(f_{kl}) &= \mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f_{kl}(\mathbf{X}_k) < f_{kl}(\mathbf{X}_l)\big\} \\ &\quad + \frac{1}{2}\mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f_{kl}(\mathbf{X}_k) = f_{kl}(\mathbf{X}_l)\big\}\,. \end{aligned}$$

For a single ranking function $f : \mathcal{X} \to \mathbb{R}$, the pairwise expected ranking accuracy is defined as

$$\begin{aligned} A_{kl}(f) &= \mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f(\mathbf{X}_k) < f(\mathbf{X}_l)\big\} \\ &\quad + \frac{1}{2}\mathbf{Pr}_{\mathbf{X}_k \sim \mathcal{D}_k, \mathbf{X}_l \sim \mathcal{D}_l}\big\{f(\mathbf{X}_k) = f(\mathbf{X}_l)\big\}\,. \end{aligned}$$

Here $\mathbf{X} \sim \mathcal{D}$ denotes that random vector $\mathbf{X}$ has distribution $\mathcal{D}$. Thus, the quality of the model is in essence evaluated by looking at the probability of correctly ranked couples $(\mathbf{X}_k, \mathbf{X}_l)$ of random vectors[1]. As in this definition, we will further always associate a single random vector $\mathbf{X}_j$ with each class, and without loss of generality, we may assume that these random vectors are independently sampled according to (different) unknown distributions, in which each distribution $\mathcal{D}_j$ corresponds to the data of one particular class. These unknown conditional distributions represent the probability of observing a certain input vector, given the class label of that input vector.

From a machine learning point of view, the primary concern is not to know the pairwise relationship of classes on a finite training set (represented by the empirical distribution, observed from a finite data sample). Rather, we want to find the relationship among the unknown underlying distributions $\mathcal{D}_j$, or in other words, the relationship between classes in input space. The r conditional class distributions $\mathcal{D}_j$, represented by random vectors $\mathbf{X}_j$, generate for each of the bipartite ranking functions f_{kl} two univariate distributions of prediction scores; for any two classes $\mathcal{C}_k$ and $\mathcal{C}_l$, two random variables $f_{kl}(\mathbf{X}_k)$ and $f_{kl}(\mathbf{X}_l)$ can be distinguished. In essence, we investigate whether the distributions $\mathcal{D}_j$ allow for an overall representation of these pairwise prediction score distributions as if they resulted from a single ranking function. Remark that the relationship between classes may not be interpreted here as a statistical dependence between classes, because data from different classes is of course independently sampled, and as such, the random vectors $\mathbf{X}_j$ are independent. We rather allude with the term relationship to the localization of the distributions in input space.

It is important to note that we will not require that the distributions of prediction scores generated by a single ranking function have to be identical to those generated by a set of bipartite ranking functions, since that would give too strong a condition. We will only enforce that the pairs of prediction score distributions have the same level of separability for both types of models, i.e. we require that the same pairwise expected ranking accuracies are obtained with a set of bipartite ranking functions and a single ranking function.

For two classes $\mathcal{C}_k$ and $\mathcal{C}_l$, the expected ranking accuracy can be expressed in terms of the joint cumulative distribution function $F_{\mathbf{X}_k, \mathbf{X}_l}$ of the random vectors $\mathbf{X}_k$ and $\mathbf{X}_l$:

$$\begin{aligned} A_{kl}(f_{kl}) &= \int_{f_{kl}(\mathbf{x}_i) < f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k, \mathbf{X}_l}(\mathbf{x}_i, \mathbf{x}_j) \\ &\quad + \frac{1}{2}\int_{f_{kl}(\mathbf{x}_i) = f_{kl}(\mathbf{x}_j)} dF_{\mathbf{X}_k, \mathbf{X}_l}(\mathbf{x}_i, \mathbf{x}_j)\,. \end{aligned}$$

As all random vectors are mutually independent, the joint cumulative distribution function of a couple can obviously be written as a product of its marginals. Given the definition of expected ranking accuracy, we introduce the concept ERA ranking representability.

Definition 3.2: Let $\mathbf{X}_1, ..., \mathbf{X}_r$ be r independent random vectors with respective conditional class distributions $\mathcal{D}_1, ..., \mathcal{D}_r$. We call a set $\overline{\mathcal{F}}$ of bipartite ranking functions ERA ranking representable on $\mathbf{X}_1, ..., \mathbf{X}_r$ if there exists a ranking function $f : \mathcal{X} \to \mathbb{R}$ such that for all $1 \leq k < l \leq r$ it holds that

$$A_{kl}(f_{kl}) = A_{kl}(f)\,. \tag{2}$$

Below we briefly discuss a way to verify ERA ranking representability. In essence, we are looking for a condition for which the set of bipartite ranking functions can be replaced by a single ranking function that gives evidence of the same expected ranking accuracy. We will see at the end that in that case the expected ranking accuracies satisfy a specific type of transitivity. This transitivity property will actually establish a condition on the distributions $\mathcal{D}_j$, but the condition itself will turn out to be distribution-independent, in the sense that the same condition must hold for any set of distributions $\mathcal{D}_1, ..., \mathcal{D}_r$. The details are given in Section IV, but we will first describe the finite sample case, for which some aspects of our story can be described in a less abstract way. Since the underlying distribution of the data is in general unknown, one obviously cannot compute the expected ranking accuracy, but one can estimate it on the basis of

[1] A restriction to vectorial input spaces is in fact not mandatory. We only make this restriction because random vector is a statistically more established concept than the more general *random data object*.

a finite labeled data sample $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$. This can be realized by computing the pairwise AUC, a nonparametric unbiased estimator of the expected ranking accuracy [11]. Thus, a ROC curve is constructed for each pair of classes. The AUC can be formally defined as follows [19]–[21].

Definition 3.3: For a set $\overline{\mathcal{F}}$ of bipartite ranking functions, we define the pairwise AUC between classes $\mathcal{C}_k$ and $\mathcal{C}_l$ for the ranking function f_{kl} with $1 \leq k < l \leq r$ as

$$\widehat{A}_{kl}(f_{kl}, D) = \frac{1}{n_k n_l} \sum_{y_i = \mathcal{C}_k} \sum_{y_j = \mathcal{C}_l} I_{f_{kl}(\mathbf{x}_i) < f_{kl}(\mathbf{x}_j)} \,. \quad (3)$$

For a single ranking function $f : \mathcal{X} \to \mathbb{R}$, the pairwise AUC is defined as

$$\widehat{A}_{kl}(f, D) = \frac{1}{n_k n_l} \sum_{y_i = \mathcal{C}_k} \sum_{y_j = \mathcal{C}_l} I_{f(\mathbf{x}_i) < f(\mathbf{x}_j)} \,.$$

Remark that I denotes the indicator function that returns one when its argument is true and zero otherwise.

For further details on this definition and a general discussion of ROC analysis in multi-class settings, we refer for example to [22]–[25]. Interestingly, it has been shown by [26]–[28] that the binary AUC is equivalent to the *Wilcoxon-Mann-Whitney* statistic. It measures the expected ranking accuracy on the empirical distribution instead of the unknown underlying distribution and, by definition, it also satisfies the reciprocity property. Given a finite data sample, the AUC allows us to define the following form of ranking representability that can be interpreted as ERA ranking representability of the observed empirical distribution.

Definition 3.4: We call a set $\overline{\mathcal{F}}$ of bipartite ranking functions AUC ranking representable on D if there exists a ranking function $f : \mathcal{X} \to \mathbb{R}$ such that for all $1 \leq k < l \leq r$ it holds that

$$\widehat{A}_{kl}(f_{kl}, D) = \widehat{A}_{kl}(f, D) \,. \quad (4)$$

In [13], we introduced AUC ranking representability as a relaxation of strict ranking representability, which basically assumes that all bipartite ranking functions must be consistent with a global ranking function. We showed that strict ranking representability can be easily verified by investigating whether a graph is free of cycles. Unfortunately, strict ranking representability has a very limited applicability, since it is a condition that cannot be satisfied for realistic data samples. However, from a statistical perspective, such a strong condition is not required and that was our main motivation to relax this condition to AUC ranking representability.

AUC ranking representability can be easily verified for small data samples by enumerating all possible rankings of the data and computing for each of them the pairwise AUCs, as shown by the example in the introduction. A more formal connection is summarized in the following.

Definition 3.5: An (a^*, s)-split, denoted $\mathbf{a}^*$, is an increasing ordered list (or vector) $\mathbf{a}^* = (a_1^*, a_2^*, ..., a_s^*)$ of s (not necessarily strictly) positive integers summing up to a^*. An (a^*, s, t)-split is an (a^*, s)-split for which each component of $\mathbf{a}^*$ is upper bounded by t. The set of all (a^*, s, t)-splits will be denoted $\mathfrak{S}(a^*, s, t)$. We define the dual $\mathbf{b}$ of an (a^*, s, t)-split as the decreasing vector $\mathbf{b}^* = (a_s^*, a_{s-1}^*, ..., a_1^*)$. The set of all dual (a^*, s, t)-splits will be denoted $\widetilde{\mathfrak{S}}(a^*, s, t)$.

Example 3.6: We give two simple examples to illustrate the above definition:

$$\begin{aligned} \mathfrak{S}(10, 4, 3) &= \{(1, 3, 3, 3), (2, 2, 3, 3)\} \,. \\ \widetilde{\mathfrak{S}}(11, 3, 6) &= \{(6, 5, 0), (6, 4, 1), (6, 3, 2), \\ &\quad (5, 5, 1), (5, 4, 2), (5, 3, 3)\} \,. \end{aligned}$$

Definition 3.7: Let $(n_1, ..., n_r) \in \mathbb{N}^r$ and let

$$\mho_{kl} = \{a \in [0, 1] \mid (\exists a^* \in \mathbb{N})(a = \frac{a^*}{n_k n_l})\} \,.$$

The family of functions $C_{jkl} : \mho_{jk} \times \mho_{kl} \to \mho_{jl}$ is defined by:

$$C_{jkl}(a, b) = \frac{1}{n_j n_l} \min_{\substack{\mathbf{a}^* \in \mathfrak{S}(a^*, n_k, n_j) \\ \mathbf{b}^* \in \widetilde{\mathfrak{S}}(b^*, n_k, n_l)}} \sum_{i=1}^{n_k} (a_i^* - a_{i-1}^*) b_i^* \,,$$

for $j, k, l \in \{1, ..., r\}$.

The value $C_{jkl}(a, b)$ is the solution of an integer quadratic program. To illustrate this, let us rewrite the minimization as:

$$\min_{\mathbf{a}^*, \mathbf{b}^*} \frac{1}{n_j n_l} \sum_{i=1}^{n_k} (a_i^* - a_{i-1}^*) b_i^*$$

$$\text{subject to} \begin{cases} \sum_{i=1}^{n_k} a_i^* = a^* \,, \\ \sum_{i=1}^{n_k} b_i^* = b^* \,, \\ a_i^* \geq a_{i-1}^* \,, \forall i \in \{1, ..., n_k\} \,, \\ b_i^* \leq b_{i-1}^* \,, \forall i \in \{2, ..., n_k + 1\} \,, \\ 0 \leq a_i^* \leq n_j \,, \forall i \in \{1, ..., n_k\} \,, \\ 0 \leq b_i^* \leq n_l \,, \forall i \in \{1, ..., n_k\} \,, \\ a_i^*, b_i^* \in \mathbb{N} \,, \forall i \in \{1, ..., n_k\} \,, \\ a_0^* = 0 \,, b_{n_k+1}^* = 0 \,. \end{cases} \quad (5)$$

Based on this definition, let us introduce a new type of transitivity.

Definition 3.8: A reciprocal relation of pairwise AUCs $\widehat{A}_{kl}(f_{kl}, D)$ is called AUC transitive if for all $j, k, l \in \{1, ..., r\}$ it holds that

$$C_{jkl}(\widehat{A}_{jk}, \widehat{A}_{kl}) \leq \widehat{A}_{jl} \,. \quad (6)$$

We emphasize that this type of transitivity in certain sense differs from all existing types of transitivity, since the condition that a given triplet of values must satisfy depends on their indices.

Theorem 3.9: A triplet $\overline{\mathcal{F}} = \{f_{12}, f_{23}, f_{13}\}$ of bipartite ranking functions is AUC ranking representable on D if and only if the corresponding reciprocal relation of AUCs is AUC transitive.

IV. ERA ranking representability

Since AUC transitivity acts as a necessary and sufficient condition for AUC ranking representability, it is able to reveal deeper insights of multi-class classifiers, but it is not of great practical value. An integer quadratic program needs to be solved, which is an NP-hard problem [29], and as a result, the condition can only be exactly verified for small data sets. Instead of focussing on intelligent algorithms to solve the integer quadratic program approximately, we will present another approach to circumvent this computational bottleneck. Simultaneously, an analytical expression for the solution of the integer quadratic program is derived.

Using the concepts from the previous section, ERA ranking representability naturally follows from AUC ranking representability by considering the abstraction from a finite sample to the underlying distribution. Let us now introduce a specific type of C-transitivity with C a conjunctor.

Definition 4.1: A reciprocal relation $Q : \mathcal{X}^2 \to [0, 1]$ is called ERA-transitive if it is C-transitive w.r.t. the conjunctor $C_{\mathbf{P}0}$ defined by

$$C_{\mathbf{P}0}(a, b) \quad = \quad \begin{cases} 0, & \text{if } a + b \leq 1\,, \\ ab, & \text{if } a + b > 1\,. \end{cases}$$

We refer to [30] for definitions of C-transitivity, stochastic transitivity and the general umbrella of cycle transitivity. Remarkably, we can show that ERA transitivity leads to a necessary and sufficient condition for ERA ranking representability.

Proposition 4.2: A triplet $\overline{\mathcal{F}} = \{f_{12}, f_{23}, f_{13}\}$ of bipartite ranking functions is ERA ranking representable on three independent random vectors if and only if the corresponding reciprocal relation of expected ranking accuracies is ERA-transitive.

Proposition 4.3: ERA transitivity implies moderate product transitivity and therefore also dice transitivity.

These propositions mainly confirm that all pieces of the puzzle fit surprisingly well. In the previous sections it was shown how AUC transitivity induces a sufficient condition for AUC ranking representability, while dice transitivity could only lead to a necessary condition. From this we were able to prove indirectly that the former type of transitivity had to be stronger than the latter one, but this could not be observed directly from the upper bound functions. Since this relationship between both types of cycle transitivity can be observed very easily in the infinite case, it gives an additional confirmation of the correctness of our analysis in the finite case.

V. Conclusion

From a machine learning point of view, we investigated whether a pairwise multi-class classification model can be simplified to a ranking model (an ordinal regression model to be more precise). To this end, we started from the assumption that the optimal complexity of a multi-class classifier is problem specific (data dependent). Reducing a pairwise multi-class classifier to an ordinal regression model can be seen as a quite drastic application of the bias-variance trade-off: a pairwise multi-class classifier is complex, containing many parameters that result in a low bias and a high variance of the performance, while an ordinal regression model contains substantially less parameters, leading to a high bias, but a low variance. So, we did not claim that a pairwise multi-class classifier can always be reduced to an ordinal regression model, we rather looked for necessary and sufficient conditions that allow for such a reduction, by analyzing the pairwise expected ranking accuracies. The result that we obtained is in this regard remarkable and important, as it confirms that the optimal complexity of a multi-class classification model depends on the distribution of the data. The conditions that we derived are moreover distribution independent, meaning that they hold for any distribution of the data.

Acknowledgment

W.W. is supported as a postdoc by the Research Foundation of Flanders (FWO-Vlaanderen).

References

[1] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The Annals of Statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[2] J. Fürnkranz, "Round robin classification," *Journal of Machine Learning Research*, vol. 2, pp. 723–747, 2002.

[3] R. Rifkin and A. Klautau, "In defense of one-versus-all classification," *Journal of Machine Learning Research*, vol. 5, pp. 101–143, 2004.

[4] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society, Series B*, vol. 42, no. 2, pp. 109–142, 1980.

[5] A. Agresti, *Categorical Data Analysis, 2nd version*. John Wiley and Sons, 2002.

[6] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems 16, Vancouver, Canada*. MIT Press, 2003, pp. 937–944.

[7] W. Chu and S. Keerthi, "Support vector ordinal regression," *Neural Computation*, vol. 19, no. 3, pp. 792–815, 2007.

[8] V. Torra, J. Domingo-Ferrer, J. Mateo-Sanz, and M. Ng, "Regression for ordinal variables without underlying continuous variables," *Information Sciences*, vol. 176, pp. 465–476, 2006.

[9] E. Hüllermeier and J. Hühn, "Is an ordinal class structure useful in classifier learning?" *International Journal of Data Mining, Modelling and Management*, vol. 1, no. 1, pp. 45–67, 2009.

[10] J. Fürnkranz, E. Hüllermeier, and S. Vanderlooy, "Binary decomposition methods for multipartite ranking," *Lecture Notes in Computer Science*, vol. 5781, pp. 359–374, 2009.

[11] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, "Generalization bounds for the area under the ROC curve," *Journal of Machine Learning Research*, vol. 6, pp. 393–425, 2005.

[12] W. Waegeman, B. De Baets, and L. Boullart, "ROC analysis in ordinal regression learning," *Pattern Recognition Letters*, vol. 29, pp. 1–9, 2008.

[13] W. Waegeman and B. De Baets, "On the ERA ranking representability of multi-class classifiers," *Artificial Intelligence*, vol. 175, pp. 1223–1250, 2011.

[14] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. MIT Press, 2000, pp. 115–132.

[15] K. Crammer and Y. Singer, "Pranking with ranking," in *Proceedings of the Conference on Neural Information Processing Systems, Vancouver, Canada*, 2001, pp. 641–647.

[16] E. Hüllermeier and J. Fürnkranz, "Pairwise preference learning and ranking," in *Proceedings of the European Conference on Machine Learning, Dubrovnik, Croatia*, 2003, pp. 145–156.

[17] S. Clémençon and N. Vayatis, "Ranking the best instances," *Journal of Machine Learning Research*, vol. 8, pp. 2671–2699, 2007.

[18] F. Wu, C. Lin, and R. Weng, "Probability estimates for multi-class support vector machines by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[19] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, pp. 203–231, 2001.

[20] P. Flach, "The geometry of ROC space: Understanding machine learning metrics through ROC isometrics," in *Proceedings of the International Conference on Machine Learning, Washington, D.C., USA*, 2003.

[21] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[22] D. Hand and R. Till, "A simple generalization of the area under the ROC curve for multiple class problems," *Machine Learning*, vol. 45, pp. 171–186, 2001.

[23] C. Ferri, J. Hernandez-Orallo, and M. Salido, "Volume under ROC surface for multi-class problems." in *Proceedings of the European Conference on Machine Learning, Dubrovnik, Croatia*, 2003, pp. 108–120.

[24] P. Flach, "The many faces of ROC analysis in machine learning," Tutorial presented at the European Conference on Machine Learning, Valencia, Spain, August 2004.

[25] J. Fieldsend and M. Everson, "Formulation and comparison of multi-class ROC surfaces." in *Proceedings of the ICML Workshop on ROC Analysis in Machine Learning, Bonn, Germany*, 2005, pp. 49–56.

[26] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristics curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[27] C. Cortes and M. Mohri, "AUC optimization versus error rate minimization," in *Advances in Neural Information Processing Systems 16, Vancouver, Canada*. MIT Press, 2003, pp. 313–320.

[28] L. Yan, R. Dodier, M. Mozer, and R. Wolniewiecz, "Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic," in *Proceedings of the International Conference on Machine Learning, Washington D.C., USA*, 2003, pp. 848–855.

[29] Z. Hua, B. Zhang, and X. Xu, "A new variable reduction technique for convex integer quadratic programs," *Applied Mathematical Modelling*, vol. 32, pp. 224–231, 2008.

[30] B. De Baets, H. De Meyer, B. De Schuymer, and S. Jenei, "Cyclic evaluation of transitivity of reciprocal relations," *Social Choice and Welfare*, vol. 26, pp. 217–238, 2006.